

Small-Sample Equating With Prior Information

*Samuel A. Livingston
Charles Lewis*

June 2009

ETS RR-09-25



Small-Sample Equating With Prior Information

Samuel A. Livingston and Charles Lewis
ETS, Princeton, New Jersey

June 2009

Abstract

This report proposes an empirical Bayes approach to the problem of equating scores on test forms taken by very small numbers of test takers. The equated score is estimated separately at each score point, making it unnecessary to model either the score distribution or the equating transformation. Prior information comes from equatings of other test forms, with an appropriate adjustment for possible differences in test length.

Key words: Equating, equipercentile equating, small sample, empirical Bayes, collateral information

The Problem

Often, a new form of a test is taken at its first administration by a small number of test takers. For some tests, the number is usually smaller than 100, frequently smaller than 50, and sometimes smaller than 20. Equating scores on the basis of such a small sample of test takers is likely to produce a poor estimate of the equating relationship in the target population. Yet, the statisticians responsible for determining the raw-to-scale score conversion need an estimate of the equating relationship in time to report scores for that small group of test takers. That estimate needs to be the best estimate of the equating relationship in the target population that can be obtained with the information available. Smoothing the score distributions before equating tends to improve the accuracy of the equating (see, for example, Livingston, 1993), but if the samples are very small, the score distributions in the samples may differ from the distributions in the population in ways that smoothing will not correct.

One possibility for improving the stability and accuracy of an estimate is to incorporate collateral information into the estimation process. The use of collateral information to improve the accuracy of an estimate in educational testing dates back at least to Kelley (1947, quoted in Lord & Novick, 1968, p. 65), who proposed using information from a group of test takers in the estimation of the true score of an individual member of the group. More recently, statisticians have used collateral information to improve the accuracy of predictions based on test scores (Rubin, 1980), including at least one application involving very small samples (Braun, Jones, Rubin, & Thayer, 1983). However, we are not aware of any previous attempt to improve the form-to-form equating of test scores by using collateral information from the equating of other test forms. Those test forms can be previous forms of the same test or, alternatively, forms of other tests that are (in ways that are important to the equating process) similar to the test form to be equated. The important similarities might include the type of items, the extent to which the items are interdependent, the approximate length of the test, and (for anchor equating designs) the characteristics of the equating anchor.

Relevant prior information is often available. The question is how to incorporate it into the process of estimating the equating relationship. A practical procedure need not be theoretically optimal, but it must improve the accuracy of the estimate—if not in every case, then at least often enough to justify its use. It must be capable of being implemented fairly easily. And it should be capable of estimating a curvilinear equating relationship. Two unequally difficult

test forms, administered to the same population, tend to produce differently skewed distributions of scores, resulting in a curvilinear equating relationship. An equating method that requires the equating transformation to be linear is likely to be inaccurate in this common situation.

Point by Point

Incorporating collateral information into an estimate of the equating transformation is a complex problem. The approach proposed here is to simplify the problem by reducing it to a series of replications of a much simpler task—estimating the equated score that corresponds to a single possible score on the new form. If this task can be accomplished for every possible raw score on the new form, the result will be an estimate of the equating transformation.¹ (If the resulting estimate of the equating function is not smooth, the estimate can be improved by smoothing it.²) The point-by-point approach to equating is not a new concept; it is actually the basis for equipercentile equating (Angoff, 1984, pp. 97–101; Kolen & Brennan, 2004, pp. 36–46). In this respect, it differs not only from linear equating but also from kernel equating (von Davier, Holland, & Thayer, 2004).³

Given a possible raw score x on the new form, the corresponding equated score on the reference form can be estimated by a weighted average of a current value and a prior value. The current value is the equated score from the current equating, denoted here as $y_{current}$. The prior value is y_{prior} . The weight of $y_{current}$ should depend on the confidence it is appropriate to place in the results of the current equating. The larger the sample, the more stable those results are likely to be, and the more weight it is appropriate to give to $y_{current}$. The weight of y_{prior} should depend on how many prior equatings contribute to it, on the stability of the information those equatings provide, and on the extent to which the information they provide is consistent. All of these factors affect the stability and the usefulness of y_{prior} as a prior estimate. The more stable the prior equatings are and the more consistent they are with each other, the more weight it is appropriate to give to y_{prior} .

A solution that has these desired properties is the empirical Bayes (EB) estimate

$$\hat{y}_{EB} = \frac{\frac{1}{\text{var}(y_{current})} y_{current} + \frac{1}{\text{var}(y_{prior})} y_{prior}}{\frac{1}{\text{var}(y_{current})} + \frac{1}{\text{var}(y_{prior})}} = \frac{y_{current} \text{var}(y_{prior}) + y_{prior} \text{var}(y_{current})}{\text{var}(y_{prior}) + \text{var}(y_{current})}, \quad (1)$$

with appropriate estimates for the variances.

To adopt this approach is to treat the current equating as if it had been sampled at random from a large domain of possible equatings, each with its own new form, reference form, and samples of test takers. For each combination of new form and reference form in this domain, there is a true equated score y —the equated score that would result from averaging over many replications of the equating procedure, each with a different pair of samples of test takers. The current equating is a member of that domain. Therefore, its value of y will enter into any estimates of quantities that refer to the entire domain of possible equatings.

If there have been many previous forms of the test that is being equated, the domain of possible equatings can be restricted to include only forms of that test. Often, however, there will be few previous forms to include—possibly none at all. It will be necessary to specify the domain of possible equatings so as to include other tests that are similar to the test to be equated. In this case, using collateral information from other tests can provide a better estimate of the heterogeneity of the domain.

To obtain a value for y_{prior} and estimate its variance, consider the domain of all possible equatings that are relevant to the current equating (including the current equating itself). In each of these possible equatings, for any given new-form score x , there is a corresponding reference-form score y . Those y values form a distribution, and that distribution has a mean and a variance. If the mean of this distribution were known, it could be used as the prior value. This distribution cannot be observed, but a small sample of it can be—the values of y that correspond to x in the prior equatings identified as relevant. The mean of the values that can be observed is an estimate of the mean of the entire distribution, and that mean, denoted here as \bar{y} , can be used as the prior value. For an estimate of $\text{var}(y_{current})$, use the square of the conditional standard error of equating, as estimated by a procedure appropriate for the equating method used. For an estimate of $\text{var}(y_{prior})$, use the following expression,⁴ if its value is greater than 0 (and 0 if it is not):

$$\hat{\sigma}_t^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2 - \frac{1}{m} \sum_{i=1}^m \hat{\sigma}_i^2 . \quad (2)$$

In this notation, i indexes the equatings, with $i = m$ for the current equating and $i = 1$ to $m - 1$ for the prior equatings used as collateral information. The observed y value in equating i is denoted as y_i , and the mean of the y_i values is denoted as \bar{y} . The term $\hat{\sigma}_i^2$ indicates an estimate of the square of the conditional standard error of the i^{th} equating, at the relevant score level. (Often an estimate is available from the equating software.)

If the conditional standard error of equating at a particular score level is large (as it often is, in the tails of the score distribution), the second term of Equation 2 may well be larger than the first term. In that case, the resulting estimate of 0 for the prior variance will cause Equation 1 to simplify to $\hat{y}_{EB} = y_{\text{prior}}$. The EB estimate of the equated score will be the mean of the values observed in all the relevant equatings. This estimate includes the current equating but gives it no more weight than any of the others.

A Complication

One complication that users of this approach are likely to encounter is that the test forms in the domain of equatings can differ in length. Even if the domain is limited to forms of a single test, some forms may have one or more items excluded from scoring (because of problems with content, printing errors, etc.). Therefore, a preliminary step before applying the proposed procedure is to transform, into percentage terms, the scores on the new form and the reference form, in the current equating and in all the equatings to be included as collateral information. This transformation consists of subtracting the lowest possible score, dividing by the range of possible scores, and multiplying by 100. One consequence of this transformation will be that forms that differ in length will have different sets of possible scores. For example, a score of 50% is possible on a form with an even number of items but not on a form with an odd number of items. If the new form in the current equating has an even number of items, one of the raw scores to be equated will be 50%. But if in one of the prior equatings the new form has an odd number of items, the table of equated scores for that prior equating will not include a new-form raw score of 50%. It will be necessary to interpolate.

The Proposed Procedure

The sequence of steps in the proposed procedure is as follows, for a particular raw score x_* on the new form in the current equating:

1. Transform x_* to $xpct_*$, where $xpct_* = 100 \left(\frac{x_* - x_{\min}}{x_{\max} - x_{\min}} \right)$.
2. In the current equating, find the equated raw score $y_{current}$ corresponding to x_* .
Transform $y_{current}$ to $ypct_{current}$ and label it $ypct_m$ for use in the steps that follow..
3. In the first prior equating, find x_1 , the new-form raw-score value for that equating for which $xpct_1 = xpct_*$. In many cases, x_1 will not be a possible raw score, and it will be necessary to interpolate between two possible new-form raw scores. These possible scores can be denoted as x_{1+} and x_{1-} , chosen so that $xpct_{1+}$ and $xpct_{1-}$ are the nearest values of $xpct_1$ above and below $xpct_*$.
4. Repeat Step 3 for the prior equatings, to determine $ypct_2$, $ypct_3$, and so on.
5. Compute $ypct_{prior} = \frac{1}{m} \sum_{i=1}^m ypct_i$, the mean, over the current and prior equatings, of the $ypct$ value corresponding to x_* .
6. Use Equation 2 to estimate $\text{var}(ypct_{prior})$ from the values of $ypct_1$, $ypct_2$, $ypct_3$, and so on, and the estimates of the conditional standard error of equating. Note that these standard errors will need to be interpolated and rescaled to match the units of the corresponding $ypct_i$ values.
7. Use Equation 1 to estimate \hat{ypct}_{EB} .
8. Transform \hat{ypct}_{EB} back to the score scale of the reference form in the current equating.

If the function defined by the successive values of \hat{ypct}_{EB} for increasing values of x_* is not smooth, apply a smoothing procedure.

Specifying the Domain of Collateral Information

An important question in implementing this procedure is how broadly to define the domain of prior equatings to be used as collateral information. Should they be limited to previous forms of the test to be equated? If equatings of forms of other tests are to be included, in what ways must those tests be similar to the test to be equated? This question is one that can best be answered empirically, and the answer is likely to depend on the set of tests being considered. At a selected percent-correct score level on the new form (e.g., 60%), record the equated scores in all the prior equatings of test forms being considered for inclusion in the domain—possibly several forms of each test. Compare the variation between tests with the variation between different forms of the same test. Repeat this procedure at several selected score levels. If the results do not indicate a systematic difference between tests, it seems reasonable to conclude that prior equatings of other tests will provide useful collateral information.

In deciding which tests to consider for possible inclusion in the domain of collateral information, the most important factor would be the extent to which the test forms tend to differ in difficulty. Some tests are constructed from items that have been thoroughly pretested on representative samples of the test-taker population. Forms of those tests are likely to show only small differences in difficulty. Other tests are constructed from items for which no pretest information is available. Forms of those tests are likely to differ much more in difficulty.

Another important question is whether the domain of collateral information should include both possible equatings of each pair of forms (i.e., the equating of X to Y and the equating of Y to X). If both are included, the differences in difficulty will tend to cancel each other out in the estimation of y_{prior} in Equation 1. The prior estimate of the equating transformation will be very close to the identity, and the value of y will be very close to that of x , when both x and y are expressed as percentages. One way to think about this question is to ask, If in the past, the new forms in the domain have tended to be more difficult (or, alternatively, easier) than the reference forms they were equated to, does that tendency represent a genuine trend? If so, is it realistic to expect this trend to continue in the future? If the answer to either of these questions is no, it may be best to use the identity as the prior estimate. In that case, the main function of the collateral information will be to estimate $\text{var}(y_{prior})$.

Advantages and Limitations

Repeating the proposed procedure for any given score on the new form to be equated will produce an estimate of the equated score that would result from equating in a very large sample from the target population. This estimate should be, in most cases, a better estimate than the equated score implied by the current small-sample equating. It should also be, in most cases, a better estimate than could be obtained by disregarding the current equating and using, instead, the score implied by a previous equating or by the average of several previous equatings.

Because the proposed procedure estimates the equated score separately for each raw-score value, it does not require constraints on the form of the equating transformation. In particular, it is not constrained to produce an equating transformation with a particular slope, or even a transformation with a constant slope. It can use collateral information from equatings computed by different methods, even if some of those equatings were constrained to be linear and others were not.

One limitation of the proposed procedure, from a theoretical point of view, is that it is not symmetric with respect to the new form and reference form. Therefore, it is not, strictly speaking, an equating procedure. Instead, it is an estimation procedure—a procedure for estimating the results that would be obtained if the current equating could be performed with data from the full target population. Even though the function to be estimated is symmetric in X and Y , the best available estimate of it from small-sample data may not be symmetric in X and Y .

A more important limitation, from a practical point of view, is the difficulty of estimating $\text{var}(y_{\text{current}})$ in Equation 1. The commonly used formulas may not yield accurate results when used with small-sample data. Resampling studies may be necessary to develop a prediction formula for $\text{var}(y_{\text{current}})$ as a function of sample size, so as not to require any parameters to be estimated from the small-sample data.

The greatest drawback to the proposed procedure is that there are situations in which it can produce a result that is less accurate than the current equating. If there is reason to believe, a priori, that the current form differs in difficulty from its reference form in a way that the new forms in prior equatings did not, it would be unwise to use the procedure presented here. Such a situation could arise if the new form were deliberately constructed to be easier or harder than previous forms of the same test. It also could arise if the new form were being equated to a reference form known to be unusually easy or unusually difficult.

In conducting resampling studies to evaluate the proposed procedure, the authors' colleague Sooyeon Kim encountered a situation in which several prior forms of a test were similar in difficulty to the reference forms they were equated to, but one subsequent form was much harder than its reference form. The close agreement among the prior equatings gave the collateral information a heavy weight in the EB estimate, pulling the estimate toward an incorrect value (Kim, Livingston, & Lewis, 2008). The important practical question is, Which is the greater danger—being misled by the collateral information or being misled by an anomalous small-sample equating result? Some previous writers have taken an extreme point of view, suggesting that when the available samples for equating are smaller than a specified size, the right thing to do is to disregard the data entirely and assume the new form and reference form to be of equal difficulty at all score levels (Kolen & Brennan, 2004, pp. 289–290; Skaggs, 2005, p. 309). But there is an alternative that seems preferable: using available data to indicate how much weight to give to the small-sample equating results and how much to another estimate of the equating transformation.

References

- Angoff, W. H. (1984) *Scales, norms, and equivalent scores*. Princeton, NJ: ETS.
- Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. *Psychometrika*, 48, 171–181.
- Kim, S., Livingston, S. A., & Lewis, C. (2008). *Investigating the effectiveness of collateral information on small-sample equating* (ETS Research Rep. No. RR-08-52). Princeton, NJ: ETS.
- Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics*, 9, 25–44.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.
- Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, 30(1), 23–29.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, 75, 801–827.
- Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, 42, 309–330.
- van der Linden, W. J. (2006). Equating error in observed-score equating. *Applied Psychological Measurement*, 30, 355–378.
- von Davier A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Notes

- ¹ If the number of possible scores on the new form is very large, it may be necessary to select a subset of the possible raw scores, compute the equated scores for those selected raw scores, and interpolate for the raw scores not selected.
- ² See, for example, the method described by Kolen (1984). This technique is typically referred to as *postsMOOTHING* to distinguish it from the *presMOOTHING* of the score distributions before equating.
- ³ It is also quite different from the approach taken by van der Linden (2006) in which “... unlike equipercentile equating, ... there is a different function for each test taker” (p. 359).
- ⁴ The derivation is shown in the appendix.

Appendix

Derivation of Estimates for y_{prior} and $\text{var}(y_{prior})$

Let i index the equatings to be used in estimating y_{prior} and $\text{var}(y_{prior})$, from 1 to m . Each of these equatings has its own new form and its own reference form.

Given a specified raw-score value x on the new form in equating i , let y_i represent the corresponding equated score. That is, y_i is the score on the reference form in equating i that corresponds to the specified raw score x on the new form.

Suppose it were possible to repeat, infinitely many times, the process of sampling test takers to take the new form in equating i , sampling test takers to take the reference form in equating i , and performing the equating of the new form to the reference form in those samples to observe a value of y_i . If the resulting values could then be averaged over all those replications of equating i , there would be a quantity that could reasonably be considered the true equated score for equating i , at new-form score-level x . Let that quantity be denoted by $a_i = \text{avg}(y_i | i)$, using avg to mean the average over infinitely many replications of equating i . Let $e_i = y_i - a_i$ represent the amount by which a particular observed value of y_i is higher or lower than the average value a_i . Then $\text{avg}(e_i | i) = 0$ and, because the replications of equatings i and i' are independent, for any two different equatings i and i' , $\text{avg}(e_i e_{i'} | i, i') = 0$.

Let σ_i^2 represent the variance of y_i over the many replications of the sampling and equating procedure of equating i (in other words, the square of the conditional standard error of equating i at new-form raw score x).

Since a_i does not vary over the replication process, $\text{var}(a_i | i) = 0$ and $\text{cov}(e_i, a_i | i) = 0$. Therefore, $\text{avg}(e_i^2 | i) = \text{var}(e_i | i) = \text{var}(y_i | i) = \sigma_i^2$ and, if $i = i'$, $\text{avg}(e_i e_{i'} | i, i') = \text{avg}(e_i^2 | i) = \sigma_i^2$.

Let $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$, the average y -value over the m equatings indexed by i . This quantity will be the value for y_{prior} in Equation 1.

Let $s_a^2 = \frac{1}{m-1} \sum_{i=1}^m (a_i - \bar{a})^2$. This is the quantity to be estimated, to provide a value for $\text{var}(y_{\text{prior}})$ in Equation 1.

One plausible estimate of s_a^2 is $\frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2$. However, this estimate may require a correction for bias. To determine the correction needed, first rewrite the sum of squares:

$$\begin{aligned} \sum_{i=1}^m (y_i - \bar{y})^2 &= \sum_{i=1}^m [(y_i - a_i) + (a_i - \bar{a}) + (\bar{a} - \bar{y})]^2 \\ &= \sum_{i=1}^m (y_i - a_i)^2 + \sum_{i=1}^m (a_i - \bar{a})^2 + \sum_{i=1}^m (\bar{a} - \bar{y})^2 \\ &\quad + 2 \sum_{i=1}^m (y_i - a_i)(a_i - \bar{a}) + 2(\bar{a} - \bar{y}) \sum_{i=1}^m (y_i - a_i) + 2(\bar{a} - \bar{y}) \sum_{i=1}^m (a_i - \bar{a}). \end{aligned}$$

Now $y_i - a_i = e_i$ for all i , and therefore $\bar{y} - \bar{a} = \bar{e}$.

Consequently, the above expression can be rewritten as

$$\begin{aligned} \sum_{i=1}^m e_i^2 + \sum_{i=1}^m (a_i - \bar{a})^2 + m(\bar{e})^2 \\ + 2 \sum_{i=1}^m (e_i)(a_i - \bar{a}) + 2(\bar{e}) \sum_{i=1}^m (e_i) + 2(\bar{e}) \sum_{i=1}^m (a_i - \bar{a}). \end{aligned}$$

Since

$\bar{e} = \frac{1}{m} \sum_{i=1}^m e_i$, this expression can be rewritten again as

$$\begin{aligned} \sum_{i=1}^m e_i^2 + \sum_{i=1}^m (a_i - \bar{a})^2 + m \left(\frac{1}{m} \sum_{i=1}^m e_i \right)^2 \\ + 2 \sum_{i=1}^m (e_i)(a_i - \bar{a}) - 2 \left(\frac{1}{m} \sum_{i=1}^m e_i \right) \sum_{i=1}^m (e_i) - 2 \left(\frac{1}{m} \sum_{i=1}^m e_i \right) \sum_{i=1}^m (a_i - \bar{a}), \end{aligned}$$

and again, as

$$\begin{aligned} & \sum_{i=1}^m e_i^2 + \sum_{i=1}^m (a_i - \bar{a})^2 + \frac{1}{m} \sum_{i=1}^m \sum_{i'=1}^m (e_i e_{i'}) \\ & + 2 \sum_{i=1}^m (e_i)(a_i - \bar{a}) - 2 \frac{1}{m} \sum_{i=1}^m \sum_{i'=1}^m (e_i e_{i'}) - 2 \frac{1}{m} \left(\sum_{i=1}^m e_i \right) \sum_{i=1}^m (a_i - \bar{a}). \end{aligned}$$

Now the average over replications for each of the m equations can be taken. (Note that a_i does not vary over replications of equation i .) This average is:

$$\begin{aligned} & \sum_{i=1}^m \text{avg}(e_i^2 | i) + \sum_{i=1}^m (a_i - \bar{a})^2 + \frac{1}{m} \sum_{i=1}^m \sum_{i'=1}^m \text{avg}(e_i e_{i'} | i, i') \\ & + 2 \sum_{i=1}^m \text{avg}(e_i | i)(a_i - \bar{a}) - 2 \frac{1}{m} \sum_{i=1}^m \sum_{i'=1}^m \text{avg}(e_i e_{i'} | i, i') - 2 \frac{1}{m} \left(\sum_{i=1}^m \text{avg}(e_i | i) \right) \sum_{i=1}^m (a_i - \bar{a}). \end{aligned}$$

To simplify this expression further, use the following results developed earlier:

$\text{avg}(e_i | i) = 0$, $\text{avg}(e_i^2 | i) = \sigma_i^2$, $\text{avg}(e_i e_{i'} | i, i') = 0$ for $i \neq i'$, and $\text{avg}(e_i e_{i'} | i, i') = \text{avg}(e_i^2 | i) = \sigma_i^2$ for $i = i'$.

These results make the expression equal to

$$\begin{aligned} & \sum_{i=1}^m \sigma_i^2 + \sum_{i=1}^m (a_i - \bar{a})^2 + \frac{1}{m} \sum_{i=1}^m \sigma_i^2 \\ & + 2 \sum_{i=1}^m 0 \cdot (a_i - \bar{a}) - 2 \frac{1}{m} \sum_{i=1}^m \sigma_i^2 - 2 \frac{1}{m} \left(\sum_{i=1}^m 0 \right) \sum_{i=1}^m (a_i - \bar{a}) \\ & = \left(1 + \frac{1}{m} - 2 \frac{1}{m} \right) \sum_{i=1}^m \sigma_i^2 + \sum_{i=1}^m (a_i - \bar{a})^2 \\ & = \left(\frac{m-1}{m} \right) \sum_{i=1}^m \sigma_i^2 + \sum_{i=1}^m (a_i - \bar{a})^2. \end{aligned}$$

This quantity is the expectation of $\sum_{i=1}^m (y_i - \bar{y})^2$ over replications of all the m equations.

Therefore, the expectation of $\frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2$ is

$$\frac{1}{m} \sum_{i=1}^m \sigma_i^2 + \frac{1}{m-1} \sum_{i=1}^m (a_i - \bar{a})^2.$$

Suppose there are unbiased estimates $\hat{\sigma}_i^2$ for each of the σ_i^2 . Then, over replications of all the m equations ($i = 1$ to m), the expected value of

$$\frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2 - \frac{1}{m} \sum_{i=1}^m \hat{\sigma}_i^2$$

is equal to

$$\frac{1}{m-1} \sum_{i=1}^m (a_i - \bar{a})^2 = s_a^2.$$

Thus an unbiased estimate for s_a^2 is

$$\hat{s}_a^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2 - \frac{1}{m} \sum_{i=1}^m \hat{\sigma}_i^2.$$